1-1-2014

Food image classification using local appearance and global structural information

Duc Thanh Nguyen
University of Wollongong, dtn156@uow.edu.au

Zhimin Zong
University of Wollongong, zmz225@uowmail.edu.au

Philip O. Ogunbona
University of Wollongong, philipo@uow.edu.au

Yasmine Probst
University of Wollongong, yasmine@uow.edu.au

Wanqing Li
University of Wollongong, wanqing@uow.edu.au

Food image classification using local appearance and global structural information

Abstract

This paper proposes food image classification methods exploiting both local appearance and global structural information of food objects. The contribution of the paper is threefold. First, non-redundant local binary pattern (NRLBP) is used to describe the local appearance information of food objects. Second, the structural information of food objects is represented by the spatial relationship between interest points and encoded using a shape context descriptor formed from those interest points. Third, we propose two methods of integrating appearance and structural information for the description and classification of food images. We evaluated the proposed methods on two datasets. Experimental results verified that the combination of local appearance and structural features can improve classification performance.

Keywords

Food image classification, Local binary pattern, Non-redundant local binary pattern, Shape context

Disciplines

Medicine and Health Sciences | Social and Behavioral Sciences

Publication Details

Nguyen, D., Zong, z., Ogunbona, P. O., Probst, Y. & Li, W. (2014). Food image classification using local appearance and global structural information. Neurocomputing, 140 242-251.

Food Image Classification Using Local Appearance and Global Structural Information

Duc Thanh Nguyen[a], Zhimin Zong[a], Philip O. Ogunbona[a], Yasmine Probst[b], Wanqing Li[a,*]

[a]*School of Computer Science and Software Engineering,*
[b]*School of Health Sciences,*
University of Wollongong, NSW 2522
Australia

Abstract

This paper proposes food image classification methods exploiting both local appearance and global structural information of food objects. The contribution of the paper is threefold. First, non-redundant local binary pattern (NRLBP) is used to describe the local appearance information of food objects. Second, the structural information of food objects is represented by the spatial relationship between interest points and encoded using shape context descriptor formed from those interest points. Third, we propose two methods of integrating appearance and structural information for the description and classification of food images. We evaluated the proposed methods on two datasets. Experimental results verified that the combination of local appearance and structural features could contribute to better classification performance.

Keywords:
Food image classification, local binary pattern, non-redundant local binary pattern, shape context

1. Introduction

The high incidence of obesity has been linked to the nutrition and an imbalanced food intake of the human population [1]. It is believed that a better understanding of the aetiology and effective health management programs could be developed through better reporting of food intake. Conventionally, this has been achieved manually through self-reporting or recording from observation. However, numerous studies have revealed that data obtained by these means seriously underestimates food intake, and thus does not accurately reflect the habitual eating behaviour of humans in real life [2, 3, 4].

Recently, image processing and pattern recognition techniques have been applied to improve the accuracy and efficiency of food intake reporting through automatic image-based food recognition systems [5]. In these systems, a comprehensive nutrition database is used to generate a daily food intake report for individuals based on computerized recognition of food images. Motivated by the importance of the health related issues associated with food intake and the progress made to date in the application of pattern recognition-based methods, this paper focuses on developing a food image recognition and classification method. Such a recognition and classification tool forms the core of a computerized food intake reporting system. We note that the problem of food recognition is not a simple test case of object recognition and

*Corresponding Author
Email addresses: dtn156@uowmail.edu.au (Duc Thanh Nguyen), philipo@uow.edu.au (Philip O. Ogunbona), yasmine@uow.edu.au (Yasmine Probst), wanqing@uow.edu.au (Wanqing Li)

this has been observed in a number of food recognition and classification research publications [6, 7, 8, 4]. Largely, this is because of the possible variations in appearance (colour, texture, and shape) and viewpoints of food images. The problem is also exacerbated by the complexity of the recording environment, e.g. cluttered background, the presence of other objects, uncontrolled photographing conditions, and illumination conditions.

Generally speaking, the state-of-the-art methods for food image recognition and classification have used descriptors that mainly exploit appearance-based features including colour [6], texture [7] and shape [8, 4] in describing food objects. While several of these appearance-based descriptors have been successful in existing food image recognition and classification methods, structural information of food objects has been ignored. It would seem that structural information is as important as appearance information. Moreover, a combination of appearance-based features and structural features would enhance the recognition performance. In this paper, we propose to combine both local appearance and global structure in the description and classification of food images. The contributions of this paper are summarised as follows:

- To take advantage of texture as a discriminative feature in describing the appearance information of food objects, we propose the use of non-redundant local binary pattern (NRLBP) to encode the local textures of food images.
- In order to describe the structural information of food objects, we use the scale-invariant interest points [9] and employ shape context descriptor [10] to encode the spatial relationship between interest points.
- We propose two different methods to integrate both the local appearance and global structural information in describing and classifying food images.

We evaluated the proposed methods on two different datasets: the Pittsburgh Fast-Food Image (PFI) dataset [6] and our collected dataset. Experimental results showed that combining both the local appearance and global structural information could enhance the classification accuracy and outperform the baseline experiments [6] provided on the PFI dataset.

The rest of this paper is organised as follows. In Section 2 we provide a brief review of existing work in the domain of food image classification. Section 3 presents basic elements such as appearance-based features and structural features as well as how to combine those features in describing and classifying food images. Experimental results along with comparative analysis are presented in Section 4. Section 5 concludes the paper and discusses future work.

2. Related Work

In food image classification, colour has been considered as one of the important features. For example, Chen et al. [6] employed a $4 \times 4 \times 4$-bin RGB colour histogram (each bin corresponds to one of the components Red, Green, and Blue) to describe food images. Each pixel in the food image was then mapped to its closest bin in the histogram to generate a 64-dimensional feature vector representing that food image. The 64-dimensional feature vectors of all training food images were used to train a support vector machine (SVM) classifier for food image classification.

For texture information, Gabor texture features extracted on local regions of 3×3 and 4×4 and at various scales and orientations were employed in the work of Joutou and Yanai [7]. Similar to the

colour histogram, the texture features of all local regions were concatenated to create a richer and higher dimensional feature vector to describe a food image.

Shape information is also used in classification of food objects. For example, in [8], the size (counted as the area) and shape (represented by the ratio of the difference between the major and minor axes, and their sum) of bread objects were used as classification cues. However, this method requires a prior knowledge of the size and shape of food objects while such information may not always be available especially in a multi-food recognition environment with possible occlusion and variable shape.

Recently, the scale-invariant feature transform (SIFT) descriptor, introduced by David Lowe [9] and then widely used for image matching and recognition, was employed to encode the local shape of food objects. For instance, in [4], interest points and their local features were extracted using SIFT detector. The classification then proceeded frame-by-frame by matching individual SIFT features from a newly acquired food image to a database of pre-trained features. This process is similar to matching key points SIFT descriptors in [9]. The advantages of SIFT descriptors are well documented. First, SIFT features extracted at interest points are local features with high informative content. Second, they are stable under local and global perturbations in the image domain. In particular, SIFT features are invariant to image scale and rotation, and have been shown to provide robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination.

Similar to [4], but instead of encoding food images directly using SIFT features, the Bag-of-Features (BoF) approach was employed in [6, 7]. This approach is inspired by the Bag-of-Words (BoW) approach devised originally for text classification [11]. In BoF approach, codewords are represented by image features (e.g. SIFT features in this case). Each food image is then represented as a histogram of occurrence frequencies of codewords defined in a discrete vocabulary called codebook. The histogram is considered as a feature vector and used to train a discriminative classifier, e.g. SVM. In [7], colour histogram, Gabor texture features, and SIFT features were together employed to train a multiple kernel learning (MKL) SVM in which a sub-kernel was assigned to each type of features.

In general, extracting features at interest points [4, 6, 7] has advantages in capturing the local information of food images and coping with the deformation of the shape of food objects. However, to date, existing works have not considered how to exploit the spatial relationship between interest points despite the potential importance of this information for object recognition. The topology of interest points embodies the structural information of objects and thus when combined with appearance information becomes a powerful feature to discriminate an object from others.

3. Proposed Food Image Classification

In this paper, we explore combining both local appearance and global structural information in enhancing the description and classification of food images. In particular, SIFT detector [9] is used to detect interest points. Non-redundant local binary pattern (NRLBP) [12] is employed as the local textural descriptor and extracted at interest points to describe the appearance information of food objects. The topology of interest points represents the structural information of food objects and is encoded using the shape context descriptor [10]. We propose two food image classification methods to integrate the local appearance information with global structural information. The first method extends the Bag-of-Features approach [4, 6, 7] by including the structural information. The second method is based on our previous work in [13]. However, different from [13], in the second method, NRLBP is used. In the rest

of this section, we briefly describe the NRLBP (section 3.1) and shape context (section 3.2). The two classification methods then will be presented (section 3.3).

3.1. Local Binary Pattern (LBP) and Non-redundant Local binary Pattern (NR-LBP)

Local binary pattern (LBP) is an effective descriptor to describe local texture [14] and has been widely used in a range of applications including texture classification [15], object recognition [16] and detection [17, 18, 12]. The success of the LBP descriptor has been due to its robustness under illumination changes, computational simplicity and discriminative power. Figure 1 represents an example of the LBP in which the LBP code of the centre pixel (in red colour and value 20) is obtained by comparing its intensity with neighbouring pixels' intensities. The neighbouring pixels whose intensities are equal or higher than the centre pixel's are labelled as "1"; otherwise as "0".

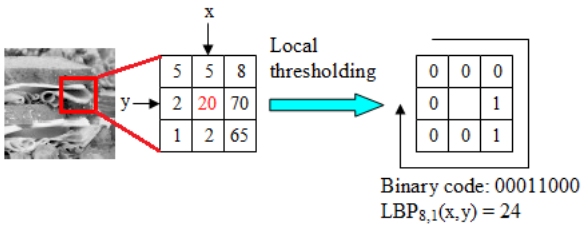


Figure 1: An illustration of the $\text{LBP}_{8,1}$ descriptor.

Consider a pixel c located at coordinate (x_c, y_c) and whose intensity is g_c. The value of the LBP code of c is defined as,

$$LBP_{M,L}(x_c, y_c) = \sum_{m=0}^{M-1} f(g_m - g_c)2^m \tag{1}$$

where m is a neighbouring pixel of c and the distance from m to c does not exceed L. The intensity of the pixel m is represented by g_m. Furthermore, the function $f(\cdot)$ is defined as,

$$f(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

In (1), M is the number of sampled points. Figure 1 shows an example of the LBP with $M = 8$ and $L = 1$. In general, the original LBP descriptor defined in (1) can generate up to 2^M different LBP codes.

An extension to the LBP is the so called uniform LBP [14] in which there are at most two bitwise transitions from 0 to 1 and vice versa in its circular binary representation. LBPs which are not uniform are called non-uniform LBPs. As indicated in [14], uniform LBPs often represent primitive structures of the texture while non-uniform LBPs usually correspond to unexpected noises and hence are less discriminative.

Recently, Nguyen et al. [12] proposed a variant of the LBP, referred to as non-redundant LBP (NRLBP), by considering a LBP code and its complement as equivalent. The NRLBP has also been applied successfully in human detection [12, 19] and smoke detection [20]. The NRLBP can be defined

as,

$$NRLBP_{M,L}(x_c, y_c) = \min\left\{LBP_{M,L}(x_c, y_c), 2^M - 1 - LBP_{M,L}(x_c, y_c)\right\} \quad (3)$$

As can be seen in (3), the NRLBP is invariant to the relative change of intensity in textural structures while it is more compact than the LBP. In this paper, both LBP and NRLBP are used to encode the appearance of the food objects.

Scanning a given image/region of image in pixel-wise fashion, LBP/NRLBP codes are accumulated into a discrete histogram called LBP/NRLBP histogram. Two given images/regions of image can be compared by measuring the similarity between their two LBP/NRLBP histograms.

3.2. Shape Context Descriptor

Shape context as a descriptor, was proposed by Belongie et al. [10] and has been widely used in various computer vision tasks, e.g. shape matching [21] and object recognition [22]. Shape matching using shape context is a variant of the conventional Hausdorff matching which employs the Hausdorff distance to measure the similarity between two shapes. The shape context descriptor can be summarised as follows.

Given a set of points P on a two-dimensional plane, the shape context at a point $p \in P$ is denoted as $s(p)$ and is the histogram of the relative coordinates of points other than p in P to p. The relative coordinate of a point $q \neq p$ to p is represented by the length (r) and orientation (θ) of the vector connecting q to p. Assuming that r and θ are quantised into $R = \{r_1, r_2, ...\}$ and $\Theta = \{\theta_1, \theta_2, ...\}$ different values respectively, $s(p)$ is a vector of $|R| \times |\Theta|$ elements, i.e. $s(p) = [s(p)_1, s(p)_2, ..., s(p)_{|R|\times|\Theta|}]$.

To make $s(p)$ more sensitive to nearby points, the histogram is represented in a log-polar space. In addition, to obtain the rotation invariance of the shape context, instead of computing the orientation θ between two points p and q in comparison to the x-axis, θ can be calculated as the angle between the vector connecting p and q and the vector connecting p to the centre location of all points in P.

3.3. Methods of Classification

The two methods of food image classification we proposed entail description and classification. For food image description, the appearance information is extracted at interest points. The structural information of a food object is embodied with respect to the location of interest points and is encoded using the shape context.

To encode the local appearance of food objects, a general codebook $\mathcal{G}$ is created as follows. Given a set of training food images, the SIFT detector [9] is first invoked to detect a set of interest points on those food images. For each interest point p, a local image patch centred at that point is extracted. The LBP/NRLBP histogram $h(p)$ of the image patch is computed and normalised using L_1 norm. All the LBP/NRLBP histograms generated on the training dataset are then clustered using a K-means algorithm in which the distance between two histograms $h(p)$ and $h(q)$ is computed using χ^2 distance,

$$d(h(p), h(q)) = \frac{1}{2}\sum_{b=1}^{B}\frac{[h(p)_b - h(q)_b]^2}{h(p)_b + h(q)_b}, \quad (4)$$

where B indicates the number of histogram bins and $h(p)_b$ is the value of $h(p)$ at the $b'th$ bin.

This step results in a general codebook $\mathcal{G} = \{w_1, w_2, ..., w_K\}$ where codewords $w_i, i \in \{1, ..., K\}$ are the LBP/NRLBP histograms nearest to the cluster means. Figure 2 shows codewords extracted on donut images. In this figure, different codewords are represented by different colours.

Figure 2: Some examples of using SIFT to locate the codewords (best viewed in colour).

3.3.1. Method I

Similar to [6, 7], this method follows the Bag-of-Features (BoF) paradigm, i.e. the descriptor of a food object is represented by the histogram of the frequency of codewords. However, in contrast to [6, 7], the spatial relationship between the codewords is exploited in our method. Figure 3 illustrates the importance of the spatial relationship; different codewords are depicted using pentagrams and circles. The left and right images of different food objects have two very different structures but the frequency of codewords (icons) in both images are identical. This observation suggests that the spatial relationship between codewords is of great importance in food image classification problem. In this method, the spatial relationship between codewords is integrated with the appearance features to form high-dimensional feature vectors describing food images.

Figure 3: Left and right images represent two different food objects with the same codeword frequency histogram. Note that the pentagrams and circles represent different codewords.

Given a trained codebook $\mathcal{G} = \{w_1, w_2, ..., w_K\}$, a descriptor is extracted from a food image I as follows. First, a set of interest points P_I on I is detected using SIFT detector [9]. For each interest point $p \in P_I$, the best matching codeword $w(p) \in \mathcal{G}$ is determined as,

$$w(p) = \arg\min_{w_i \in \mathcal{G}} d(h(w_i), h(p)), \tag{5}$$

where $d(h(w_i), h(p))$ is computed using (4).

Similar to [6, 7], the appearance frequency histogram $F = [F_1, F_2, ..., F_K]$ where K is the number of

codewords in $\mathcal{G}$ is generated so as,

$$F_i = \frac{\sum_{p \in P_I} 1(w(p) = w_i)}{\sum_{j=1}^{K} \sum_{p \in P_I} 1(w(p) = w_j)} \tag{6}$$

where $1(w(p) = w_i)$ is an indicator function, i.e., $1(w(p) = w_i) = 1$ if $w(p) = w_i$, and 0, otherwise.

The structural information of the food image I is defined based on the shape contexts of codewords determined on I as follows. For every $p \in P_I$, whose codeword is $w(p)$, its shape context, denoted as $s(p)$, is created as the histogram of the relative coordinates from p to every other point $q \in P_I$ whose $w(p) \neq w(q)$. Since every codeword $w \in \mathcal{G}$ may have more than one matched interest point, i.e. $\exists p, q \in P_I$, $p \neq q$ and $w = w(p) = w(q)$, the shape context of a codeword w is computed as the average of the shape contexts extracted at interest points whose codeword is w. That is, the shape context of a codeword $w \in \mathcal{G}$ of an image I is denoted as $s(w) = [s(w)_1, s(w)_2, ..., s(w)_{|R| \times |\Theta|}]$ in which

$$s(w)_i = \frac{\sum_{p \in P_I(w)} s(p)_i}{|P_I(w)|} \tag{7}$$

where $P_I(w)$ is a set of interest points of image I whose codeword is w and $|P_I(w)|$ denotes the cardinality of the set $P_I(w)$.

Finally, the appearance-structural feature vector V describing the food image I is formed by aggregating F and $s(w)$ as,

$$V = F \bigoplus_{w \in \mathcal{G}} s(w) \tag{8}$$

where $\bigoplus$ denotes the concatenation operation.

It can be seen from (8) that, for each food image, V has $K + K \times |R| \times |\Theta|$ elements. The feature vectors of all training images are used to train a SVM classifier which is then employed for classifying new food images.

3.3.2. Method II

This method is based on the approach presented in [13]. In the method, in addition to the general codebook $\mathcal{G}$, individual codebooks corresponding to different categories of food are also constructed. This is motivated by the observation that not all the codewords have the same level of discriminative power for each food category. Some codewords might not be sufficiently discriminative to represent a certain category of food. We take a step further from the general codebook by selecting discriminative codewords for each food category. A codeword filtering step is initiated so that we can keep typical and important codewords to characterise each food category.

Specifically, assume that there are N food categories and a set of training images $\mathcal{T}$, for each category C_i and its corresponding set of training images $\mathcal{T}_i \subset \mathcal{T}$, we compute the relative frequency of assignments

of each codeword $w \in \mathcal{G}$ to C_i, i.e. $f(w|C_i)$, and not to C_i, i.e. $f(w|NotC_i)$ as follows.

$$f(w|C_i) = \frac{\sum_{I \in \mathcal{T}_i} |P_I(w)|}{\sum_{\mathcal{T}_i \in \mathcal{T}} \sum_{I \in \mathcal{T}_i} |P_I(w)|} \tag{9}$$

$$f(w|NotC_i) = \frac{\sum_{I \in \mathcal{T} - \mathcal{T}_i} |P_I(w)|}{\sum_{\mathcal{T}_i \in \mathcal{T}} \sum_{I \in \mathcal{T}_i} |P_I(w)|} \tag{10}$$

A codeword w is selected to represent the category C_i if the following condition is satisfied:

$$\log \left[\frac{f(w|C_i)}{f(w|NotC_i)} \right] \geq \phi \tag{11}$$

where ϕ is a predefined threshold. The codeword filtering step yields a set of individual codebooks $G_1, G_2, ..., G_N$ for N different food categories. Note that for every $i, j \in \{1, ..., N\}$, $G_i \cap G_j$ may be non-empty.

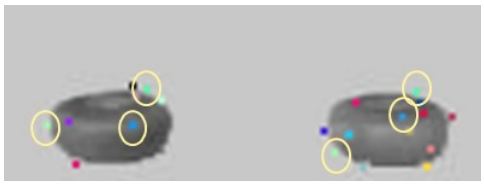


Figure 4: Different samples of a donut, each has three visual codewords marked by a circle (best viewed in colour). Note that marked codewords are the ones obtained after filtering [13].

Figure 4 shows two different samples of donut. The interest points and their codewords (for the category *donuts*) are highlighted by coloured points. Different codewords are marked by different colours. As shown in Figure 4, the spatial information of the codewords is useful for determining the correspondence between codewords.

Given a set of training images of a food category C_i, the shape context $s(w_{ij})$ of a codeword $w_{ij} \in G_i$ is computed as the average of all shape contexts of w_{ij} on all training images of the category C_i. Note that each codeword may have more than one shape context corresponding to various food categories. In addition, for each category, the shape context of a codeword is obtained based on only the codewords of the codebook for that food category.

For a test food image I, a descriptor is constructed for each category. The descriptor consists of two components describing the appearance and structural information through the set of interest points, P_I, detected from I. Specifically, every interest point $p \in P_I$ is classified into one of the words $w_{ij} \in G_i, j = 1, 2, \cdots, |G_i|$. Shape context histogram $s_I(w_{ij})$ and LBP/NRLBP histogram $h_I(w_{ij})$ for word w_{ij} are calculated from the image I and compared against their counterparts of the food category, $s(w_{ij})$ and $h(w_{ij})$. If w_{ij} does not appear in I, then both its shape context and appearance histograms are set to zero, that is, $s_I(w_{ij}) = 0$ and $h_I(w_{ij}) = 0$. The matching cost, $\mathcal{D}(I, C_i)$, between I and a food category C_i is computed as,

$$\mathcal{D}(I, C_i) = \sum_{j=1}^{|G_i|} \delta(s_I(w_{ij}), s(w_{ij})) d(h_I(w_{ij}), h(w_{ij})) \tag{12}$$

where G_i represents the set of codewords for the food category C_i, $h_I(w_{ij})$ is the mean of LBP/NRLBP

histograms extracted at interest points $p \in P_I$ whose codeword is w_{ij}, $d(h(w_{ij}), h_I(w_{ij}))$ is calculated as in (4), $\delta(s_I(w_{ij}), s(w_{ij}))$ is the χ^2-distance between two shape contexts and computed as,

$$\delta(s_I(w_{ij}), s(w_{ij})) = \frac{1}{2} \sum_{n=1}^{|R|\times|\Theta|} \frac{[s_I(w_{ij})_n - s(w_{ij})_n]^2}{s_I(w_{ij})_n + s(w_{ij})_n}, \tag{13}$$

where $s(\cdot)_n$ denotes the value of the shape context $s(\cdot)$ at the $n'th$ bin.

Note that, when only the appearance feature is used, (12) will be simplified to,

$$\mathcal{D}(I, C_i) = \sum_{j=1}^{|G_i|} d(h(w_{ij}), h_I(w_{ij})) \tag{14}$$

Finally, the classification is achieved by finding the best matching category C_i^* so that,

$$C_i^* = \arg\min_{C_i} \mathcal{D}(I, C_i), \tag{15}$$

4. Experiments and Results

4.1. Experimental Setup

In our experiments, LBP and NRLBP with $M = 8$ and $L = 1$ were employed to encode the appearance of food objects. In addition, uniform LBPs and NRLBPs were used to reduce the number of LBP/NRLBP histogram bins in which all the non-uniform LBPs/NRLBPs were cast into one bin. The dimension of the histogram was 59 for LBP and 30 for NRLBP. For the shape context descriptor, 5 and 12 bins were used respectively for r and θ; i.e. $|R| = 5$ and $|\Theta| = 12$.

The proposed methods were evaluated on two datasets: the Pittsburgh Fast-Food Image (PFI) dataset [6] and our collected datatset. PFI is the first food image dataset with a huge collection of visual data to facilitate research in automated food recognition. The dataset contains a total of 4,545 still images, 606 stereo pairs, 303 360-degree videos for structure from motion, and 27 privacy-preserving videos of eating events of volunteers. The images and videos are captured in both restaurant condition and a controlled lab setting. The food images in the PFI set are labelled as six different categories: *sandwiches*, *meat*, *salads*, *donuts*, *hamburger* and *miscellaneous*. We also collected another dataset including 343 images (43 images for training and 300 image for testing, 50 images for each category). The new dataset has 6 categories: *cakes*, *carrots*, *custards*, *milk*, *pasta*, and *pizza*. On both the datasets, food images of each category are separated into training and test sets so that no sample appears in both the sets. Some samples of the PFI dataset and our collected dataset are shown in Figure 5.

4.2. Performance Evaluation

We first evaluated the discriminative ability of the LBP and NRLBP in both method 1 and 2. The classification accuracy was used as a performance measure. Tables 1-4 show the classification accuracy of method 1 and 2 on the two datasets. As shown in those tables, in general, the NRLBP outperformed the LBP. This observation is consistent with results reported in [12]. However, the improvement was not equal for every food category in both the methods. This can be ascribed to the diversity of the texture of food objects. For example, without using structural information (column 1 and 2 in tables), on the PFI dataset, in both methods, the NRLBP contributed a significant improvement on the category *sandwiches*

(a)

(b)

Figure 5: Some samples of the PFI dataset [6] (a) and our collected dataset (b).

while retaining the same effect on the category *meat*. However, it is noteworthy that the NRLBP is more compact than the LBP.

We also verified the importance of the structural information in food image description and classification. The verification was conducted by employing only the appearance information (LBP/NRLBP) and combining the appearance information with the structural information. Note that, when only the appearance feature is used, (8) will be simplified as $V = F$ in method 1 and (14) will be used in method 2. Tables 1-4 show the results of this experiment for method 1 and 2 on the two datasets. As can be seen from those tables, in most the cases the use of structural information could improve the classification performance irrespective of the appearance features LBP/NRLBP and/or the methods used. Especially, on the PFI dataset, the improvement was significant for the category *salads* in both methods, e.g. in-

creasing by 15% when used with LBP in method 1, and 16% with NRLBP in method 1, 14% with LBP in method 2, 15% with NRLBP in method 2. On our collected dataset, method 1 shows the best advantage of structural information on the category *milk* with 24% for the use of LBP and 28% for the use of NRLBP. For method 2, the highest improvement was for the category *carrots* with 10% for both LBP and NRLBP.

Table 1: Classification accuracy of method 1 on the PFI dataset with the use of LBP, NRLBP, LBP with structural information (LBP + SI), and NRLBP with structural information (NRLBP + SI).

	LBP	NRLBP	LBP+SI	NRLBP+SI
sandwiches	0.54	0.58	0.67	0.69
meat	0.61	0.61	0.68	0.68
salads	0.75	0.78	0.90	0.91
donuts	0.46	0.47	0.48	0.48
hamburger	0.54	0.57	0.63	0.64
miscellaneous	0.67	0.67	0.72	0.72

Table 2: Classification accuracy of method 2 on the PFI dataset with the use of LBP, NRLBP, LBP with structural information (LBP + SI), and NRLBP with structural information (NRLBP + SI).

	LBP	NRLBP	LBP+SI	NRLBP+SI
sandwiches	0.79	0.81	0.82	0.86
meat	0.50	0.50	0.52	0.53
salads	0.64	0.66	0.78	0.81
donuts	0.51	0.52	0.56	0.58
hamburger	0.56	0.58	0.63	0.67
miscellaneous	0.58	0.59	0.67	0.69

Table 3: Classification accuracy of method 1 on our collected dataset with the use of LBP, NRLBP, LBP with structural information (LBP + SI), and NRLBP with structural information (NRLBP + SI).

	LBP	NRLBP	LBP+SI	NRLBP+SI
cakes	0.28	0.28	0.30	0.30
carrots	0.30	0.32	0.30	0.38
custards	0.50	0.50	0.58	0.62
milk	0.54	0.54	0.78	0.82
pasta	0.50	0.56	0.60	0.60
pizza	0.64	0.66	0.70	0.70

The classification accuracy of method 1 and 2 on the PFI dataset and our collected dataset are summarised and compared in Figure 6. Since the above experimental results have shown that the use of NRLBP and structural information could improve the classification accuracy, in those figures only the

Table 4: Classification accuracy of method 2 on our collected dataset with the use of LBP, NRLBP, LBP with structural information (LBP + SI), and NRLBP with structural information (NRLBP + SI).

	LBP	NRLBP	LBP+SI	NRLBP+SI
cakes	0.14	0.22	0.20	0.36
carrots	0.26	0.30	0.36	0.40
custards	0.46	0.60	0.52	0.70
milk	0.68	0.72	0.70	0.78
pasta	0.62	0.66	0.68	0.72
pizza	0.60	0.64	0.68	0.70

performance of the NRLBP and structural information in both the methods is presented. On the PFI dataset, for method 1, the category, *salads*, achieves the highest accuracy, while the category, *donuts*, has the lowest accuracy. On the other hand, on this set, method 2 obtains the best classification performance with the category *sandwich* and the worst performance with the category *meat*. Through experiments on the PFI dataset, we have found that the categories *donuts* and *meat* usually have smaller food items and their surfaces are less diverse. This leads to fewer interest points compared with other categories, e.g. *salads*. Therefore, on the categories *donuts* and *meat*, the shape contexts become less stable and discriminative; and thus make little improvement. On our collected data, both method 1 and 2 obtain the highest accuracy on the category *milk* with 82% for method 1 and 78% for method 2. However, both methods get poor performance on the category *cakes*. This is due to the large variation of this category in both appearance and structure compared with other food categories (as shown in Figure 5(b)).

Figure 6 also shows that method 1 and 2 achieve similar classification performance on the PFI dataset. In particular, on the categories *meat*, *salads*, and *miscellaneous*, method 1 obtains higher performance while on the categories *sandwiches*, *donuts*, and *hamburger*, method 2 shows a clear advantage. On average, the classification accuracy of method 1 and 2 is 68% and 69% respectively. On our collected data, we have found that method 2 slightly outperforms method 1 on most of the food categories (except the category *milk*). Overall, the accuracy achieved by method 1 and 2 is 57% and 61% respectively.

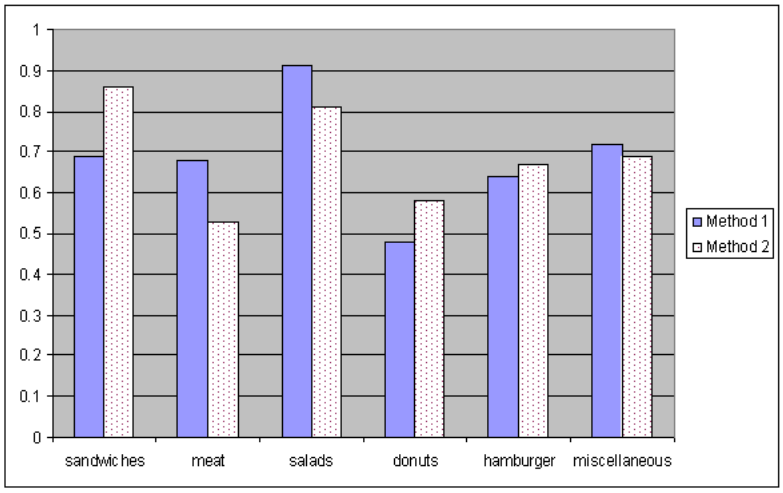
4.3. Comparison

In addition to evaluating the performance of the proposed methods, we also compared the proposed algorithms with the two baselines provided in [6].

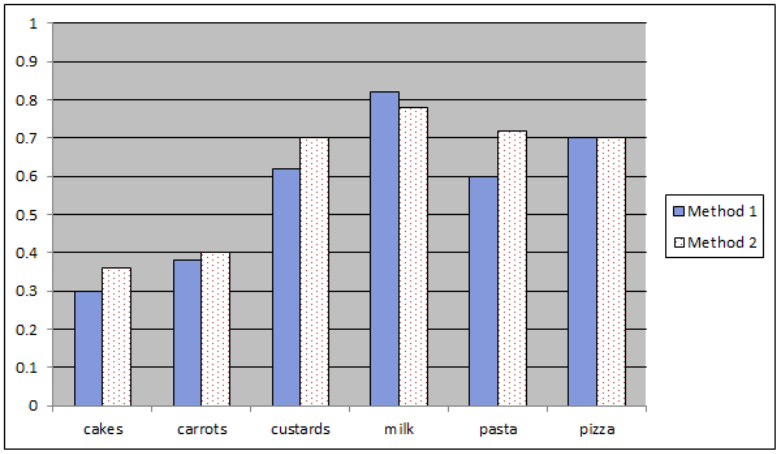
Baseline 1: Colour Histogram + SVM Classifier. In this baseline, standard RGB 3-dimensional histogram is used to describe food objects. The colour histogram is quantised with four levels per channel to form a 64-dimensional feature vector for each food image. The histogram is generated by mapping the colour of every pixel in the food image to its closest bin in the histogram. Histograms are used to train a SVM classifier which will be employed for classifying new food images.

Baseline 2: Bag of SIFT Features + SVM Classifier. In this baseline, the histogram of frequency of 128-dimensional SIFT features [9] is used as the descriptor for a food object. Similarly to baseline 1, a SVM trained by the histograms of SIFT features is used to classify food images.

Table 5 and Figure 7 represent the comparison results of method 1, 2 and the two baseline methods on the PFI dataset. Experimental results have shown that, method 1 nearly outperformed both baseline 1 and 2 on all the categories. Specifically, on the category, *salads*, which has the highest accuracy (90%)



(a)



(b)

Figure 6: Classification accuracy of method 1 and 2 on the PFI dataset (a) and our collected dataset (b).

in baseline experiments, method 1 achieved even higher accuracy (91%). On the category, *sandwiches*, the classification accuracy of method 1 was 18% higher than that of baseline 2. However, it was 5% lower than the accuracy of baseline 1. This is probably due to less the distinctive texture compared with the colour information of the category, *sandwiches*. For example, white surface is often found in *sandwiches* and thus may become important to discriminate *sandwiches* from others. On the remaining food categories, method 1 shows better performance in comparison to both the baselines.

For method 2, we have found that, the classification accuracy on the category, *salads* was 81% which was 9% lower than that of baseline 2, but 15% better than that of baseline 1. Similar situation was found in the category, *meat*, i.e. method 2 remained in the middle position compared with baseline 1 and 2. One possible reason is the variation of the texture of those categories, e.g. the category *salads* contains different ingredients with inconsistent texture and thus is not informative to characterise the category. However, the difference between the performance of all comparative algorithms was slight. For example, on the category, *meat*, method 1 achieved an accuracy of 53% while the accuracy of baseline 1 and 2 was 56% and 47% respectively. On other food categories, method 2 outperformed both the baselines.

Table 5: Comparison results of method 1, 2 and the two baselines on all food categories of the PFI dataset.

	Baseline 1	Baseline 2	Method 1	Method 2
sandwiches	0.74	0.51	0.69	0.86
meat	0.47	0.56	0.68	0.53
salads	0.66	0.90	0.91	0.81
donuts	0.21	0.43	0.48	0.58
hamburger	0.61	0.49	0.64	0.67
miscellaneous	0.31	0.64	0.72	0.69
Average accuracy	**0.50**	**0.59**	**0.68**	**0.69**

We also compared method 1 and 2 with the baseline methods based on the average classification accuracy over all food categories in the last row of Table 5. As shown in Table 5, the two proposed methods produce reasonably good results in comparison with the two baseline methods and have advantages in classifying certain categories of food. Using the average accuracy, the best performance is achieved by method 2 on the PFI dataset.

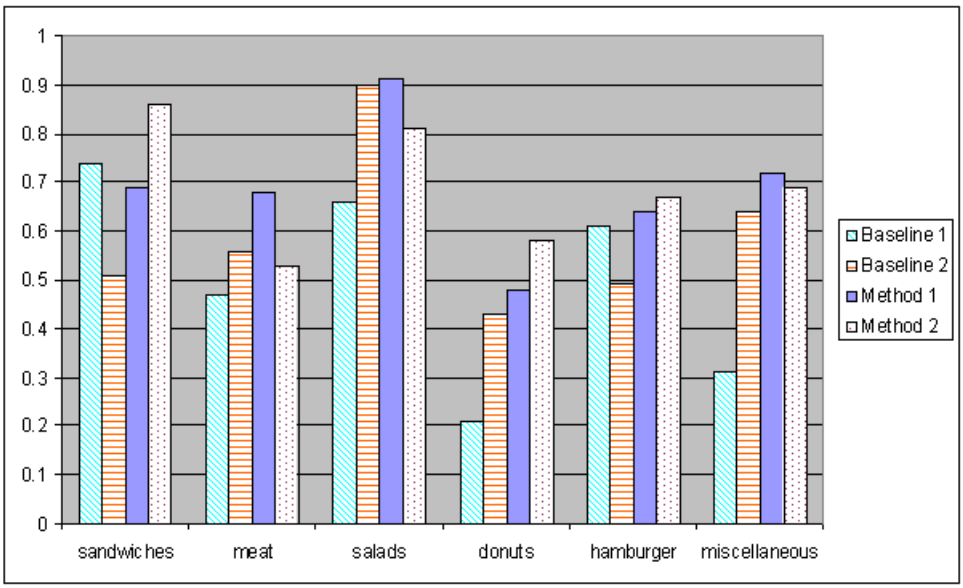


Figure 7: Comparison results of all the methods on the PFI dataset.

5. Conclusion

This paper proposes two methods for automatic classification of food images that can be used in a nutrition intake self-reporting system. Our proposed methods combine the appearance and structural information in the description and classification of food images. In particular, non-redundant local binary pattern (NRLBP) is employed as an appearance descriptor. The structural information of food objects is represented by the location of interest points and encoded using shape context. The proposed

methods were evaluated and compared with other state-of-the-art food image classification methods on two datasets. Experimental results showed the potential of employing the structural information in enhancing the description and classification of food images. Food image classification is a challenging problem due to lack of apparent visual features for some food categories and food images can be captured from various viewpoints, developing view invariant texture features would be an important future work. The possibility of combining different features to improve the classification performance would also need to be investigated. Furthermore, classification of food images containing multiple food objects and in complex background is also worthy of pursuit.

References

[1] A. Kazaks, Obesity: food intake, Primary Care: Clinics in Office Practice 30 (2) (2003) 301–316.

[2] L. Lissner, Measuring food intake in studies of obesity, Public Health Nutr. 5 (6A) (2002) 889–892.

[3] N. Yao, R. J. Sclabassi, Q. Liu, J. Yang, J. D. Fernstrom, M. H. Fernstrom, M. Sun, A video processing approach to the study of obesity, in: Proc. IEEE International Conference on Multimedia and Expo, 2007, pp. 1727–1730.

[4] W. Wu, J. Yang, Fast food recognition from videos of eating for calorie estimation, in: Proc. IEEE International Conference on Multimedia and Expo, 2009, pp. 1210–1213.

[5] L. Yang, J. Yang, N. Zheng, H. Cheng, Layered object categorization, in: Proc. IEEE International Conference on Pattern Recognition, 2008, pp. 545–576.

[6] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, J. Yang, PFID: Pittsburgh fast-food image dataset, in: Proc. IEEE International Conference on Image Processing, 2009, pp. 289–292.

[7] T. Joutou, K. Yanai, A food image recognition system with multiple kernel learning, in: Proc. IEEE International Conference on Image Processing, 2009, pp. 285–288.

[8] D. Pishva, A. Kawai, T. Shiino, Shape based segmentation and color distribution analysis with application to bread recognition, in: Proc. IAPR Workshop on Machine Vision Applications, 2000, pp. 193–196.

[9] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[10] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (4) (2002) 509–522.

[11] K. Nigam, J. Lafferty, A. Mccallum, Using maximum entropy for text classification, in: Proc. Intl. Joint. Conf. on Aritificial Intelligence Workshop on Machine Learning for Information Filtering, 1999, pp. 61–67.

[12] D. T. Nguyen, Z. Zong, P. Ogunbona, W. Li, Object detection using non-redundant local binary patterns, in: Proc. IEEE International Conference on Image Processing, 2010, pp. 4609–4612.

[13] Z. Zong, D. T. Nguyen, P. Ogunbona, W. Li, On the combination of local texture and global structure for food classification, in: Proc. IEEE International Symposium on Multimedia, 2010, pp. 204–211.

[14] T. Ojala, M. Pietikăinen, D. Harwood, A comparative study of texture measures with classification based on featured distributions, Pattern Recognition 29 (1) (1996) 51–59.

[15] T. Ojala, M. Pietikăinen, D. Harwood, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (7) (2002) 971–987.

[16] T. Ojala, M. Pietikăinen, D. Harwood, View-based recognition of real-world textures, Pattern Recognition 37 (2) (2004) 313–323.

[17] T. Ojala, M. Pietikăinen, D. Harwood, A texture-based method for modeling the background and detecting moving objects, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (4) (2006) 657–662.

[18] Y. Mu, S. Yan, Y. Liu, T. Huang, B. Zhou, Discriminative local binary patterns for human detection in personal album, in: Proc. IEEE International Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[19] D. T. Nguyen, P. Ogunbona, W. Li, Human detection with contour-based local motion binary patterns, in: Proc. IEEE International Conference on Image Processing, 2011, pp. 3609–3612.

[20] H. Tian, W. Li, P. Ogunbona, D. T. Nguyen, C. Zhan, Smoke detection in videos using non-redundant local binary pattern-based features, in: Proc. IEEE International Workshop on Multimedia Signal Processing, 2011, pp. 1–4.

[21] A. Thayananthan, B. Stenger, P. H. S. Torr, R. Cipolla, Shape context and chamfer matching in cluttered scenes, in: Proc. IEEE International Conference on Computer Vision and Pattern Recognition, 2003, pp. 127–133.

[22] G. Mori, J. Malik, Recovering 3d human body configurations using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (7) (2006) 1052–1062.